English STD 2006 Results

Jonathan Fiscus, Jérôme Ajot, George Doddington

National Institute of Standards and Technology
IAD

Spoken Term Detection Workshop
December 14-15, 2006

Outline

1. **Participants**
   - Number of Submissions
2. **Corpus Information**
   - Terms Richness
3. **Primary Results**
   - Performance - BNEWS
   - Performance - CTS
   - Performance - CONFMTG
   - Performance - ATWV Overview
   - Processing Ressources
4. **Conditional Analysis**
   - Analyses
   - Distribution
   - N-gram
   - Syllable Quantile
   - Occurrence in all sources

Outline

1. **Participants**
   - Number of Submissions
2. Corpus Information
   - Terms Richness
3. Primary Results
   - Performance - BNEWS
   - Performance - CTS
   - Performance - CONFMTG
   - Performance - ATWV Overview
   - Processing Ressources
4. Conditional Analysis
   - Analyses
   - Distribution
   - N-gram
   - Syllable Quantile
   - Occurrence in all sources

Submissions by sites and source types

Site ID	Site Name	Number of Submissions		
		BNEWS	CTS	CONFMTG
BBN	BBN Technologies		2 (1p,1c)	
BUT	Brno University of Technology	4 (1p,3c)	4 (1p,3c)	4 (1p,3c)
DOD	Department of Defense		2 (1p,1c)	
IBM	IBM	4 (1p,3c)	4 (1p,3c)	4 (1p,3c)
IDIAP	Institut Dalle Molle d'Intelligence Artificielle Perceptive	1 (1p,0c)	1 (1p,0c)	1 (1p,0c)
OGI	OGI School of Science & Technology	1 (1p,0c)	1 (1p,0c)	1 (1p,0c)
QUT	Queensland University of Technology	3 (1p,2c)	3 (1p,2c)	3 (1p,2c)
SRI	SRI International	1 (1p,0c)	1 (1p,0c)	1 (1p,0c)
STBU	Stellenbosch & Brno University Collaboration	1 (1p,0c)	1 (1p,0c)	1 (1p,0c)
STELL	Stellenbosch University	1 (1p,0c)	1 (1p,0c)	1 (1p,0c)
TUB	Technischen Universität Berlin	2 (1p,1c)	2 (1p,1c)	2 (1p,1c)
11		**18 (9p,9c)**	**22 (11p,11c)**	**18 (9p,9c)**

p: Primary submission
c: Contrastive submission

Outline

1. Participants
   - Number of Submissions
2. **Corpus Information**
   - Terms Richness
3. Primary Results
   - Performance - BNEWS
   - Performance - CTS
   - Performance - CONFMTG
   - Performance - ATWV Overview
   - Processing Ressources
4. Conditional Analysis
   - Analyses
   - Distribution
   - N-gram
   - Syllable Quantile
   - Occurrence in all sources

Occurrences

Site	# Occurences	Speech Hours	Occ./Hours
BNEWS	4893	2.212	2211.66
CTS	5856	2.993	1956.78
CONFMTG	3672	2.098	1750.06

Outline

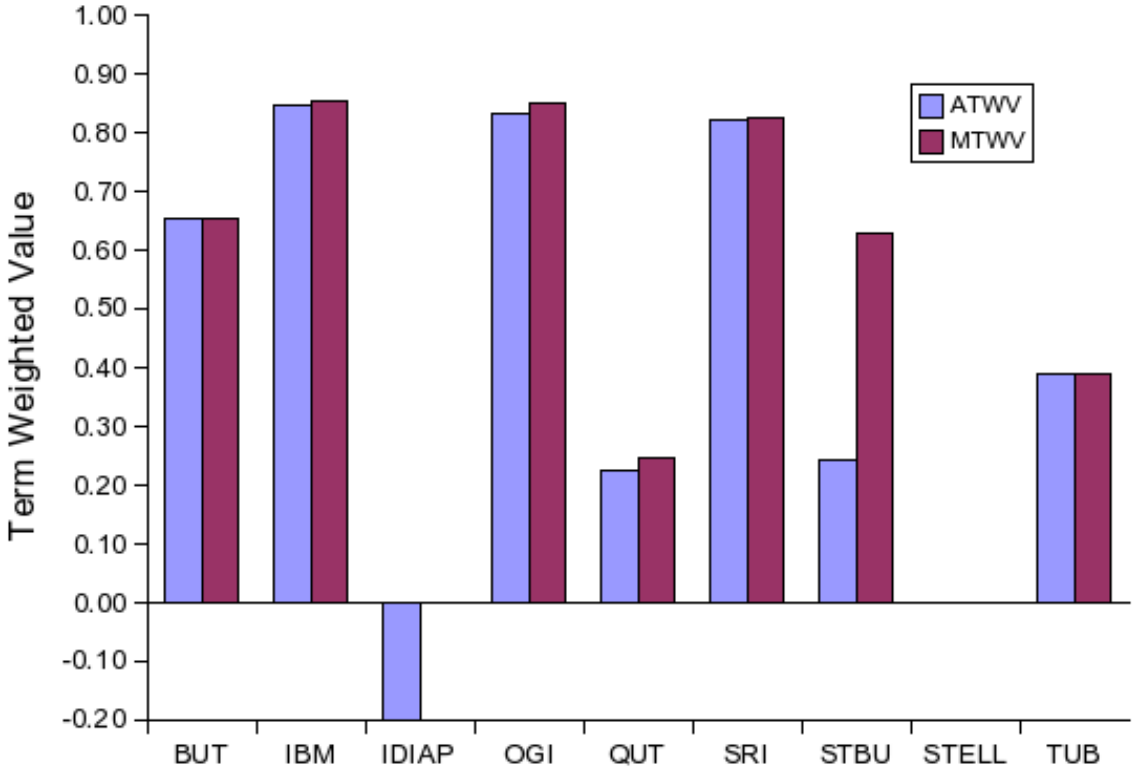
1. Participants
   - Number of Submissions
2. Corpus Information
   - Terms Richness
3. **Primary Results**
   - Performance - BNEWS
   - Performance - CTS
   - Performance - CONFMTG
   - Performance - ATWV Overview
   - Processing Ressources
4. Conditional Analysis
   - Analyses
   - Distribution
   - N-gram
   - Syllable Quantile
   - Occurrence in all sources

Performance - BNEWS

Actual and Maximum Term Weighted Value

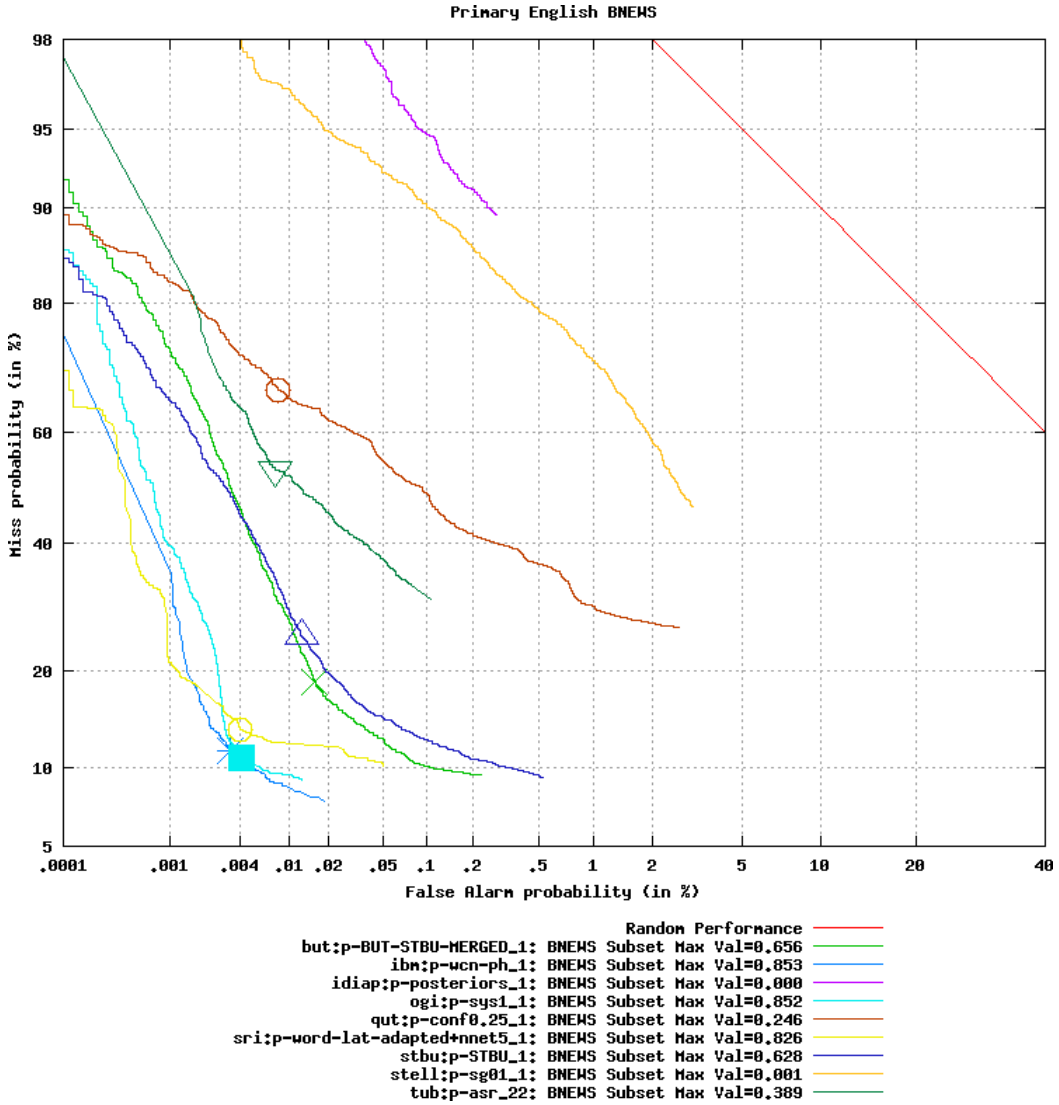
Site	ATWV	MTWV
BUT	0.6541	0.6558
IBM	0.8485	0.8532
IDIAP	-2.7139	0.0000
OGI	0.8323	0.8516
QUT	0.2265	0.2459
SRI	0.8238	0.8258
STBU	0.2434	0.6278
STELL	0.0000	0.0006
TUB	0.3890	0.3894

ATWV: Actual Term Weighted Value
MTWV: Maximum Term Weighted Value



Performance - BNEWS

DET Curve

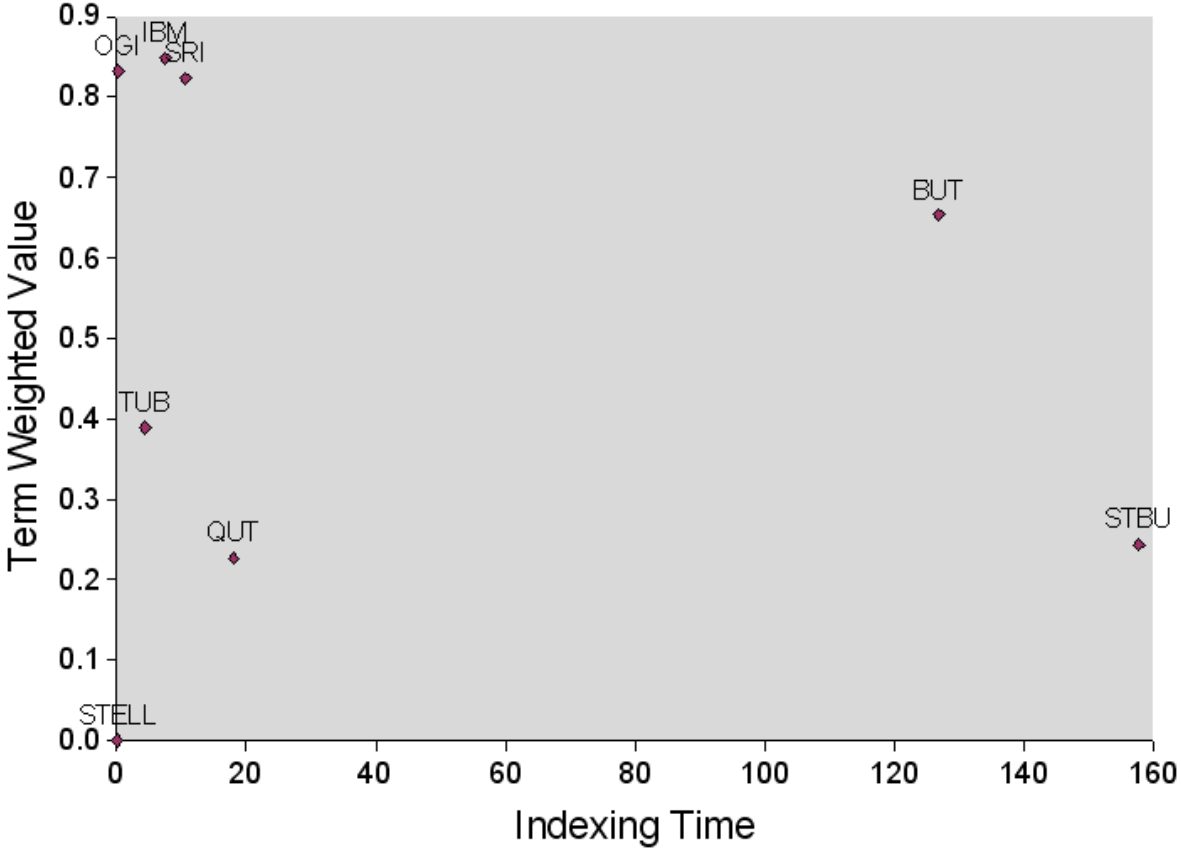


NIST National Institute of Standards and Technology

Indexing Time vs. ATWV

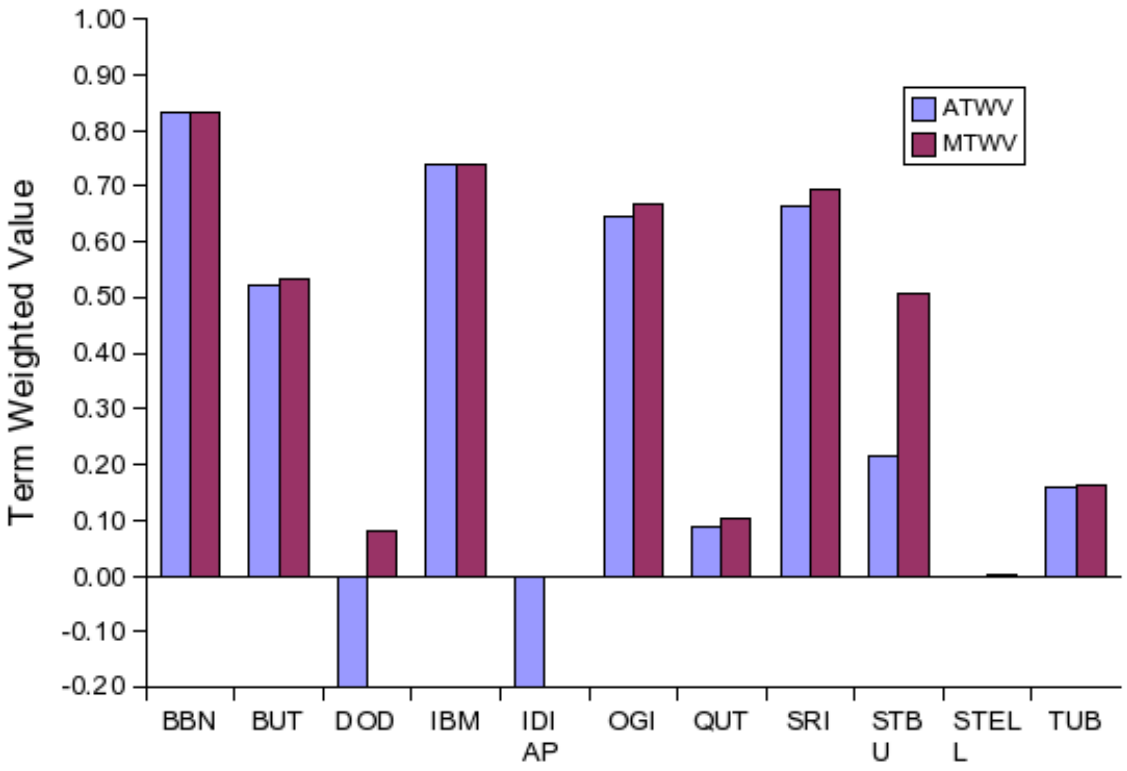
Site	Indexing Time	ATWV
BUT	126.7596	0.6541
IBM	7.5627	0.8485
IDIAP	0.2859	-2.7139
OGI	0.2891	0.8323
QUT	18.0884	0.2265
SRI	10.6555	0.8238
STBU	157.6593	0.2434
STELL	0.1680	0.0000
TUB	4.3897	0.3890

Indexing Time Unit:
Processing Hours per Hours of Speech



Actual and Maximum Term Weighted Value

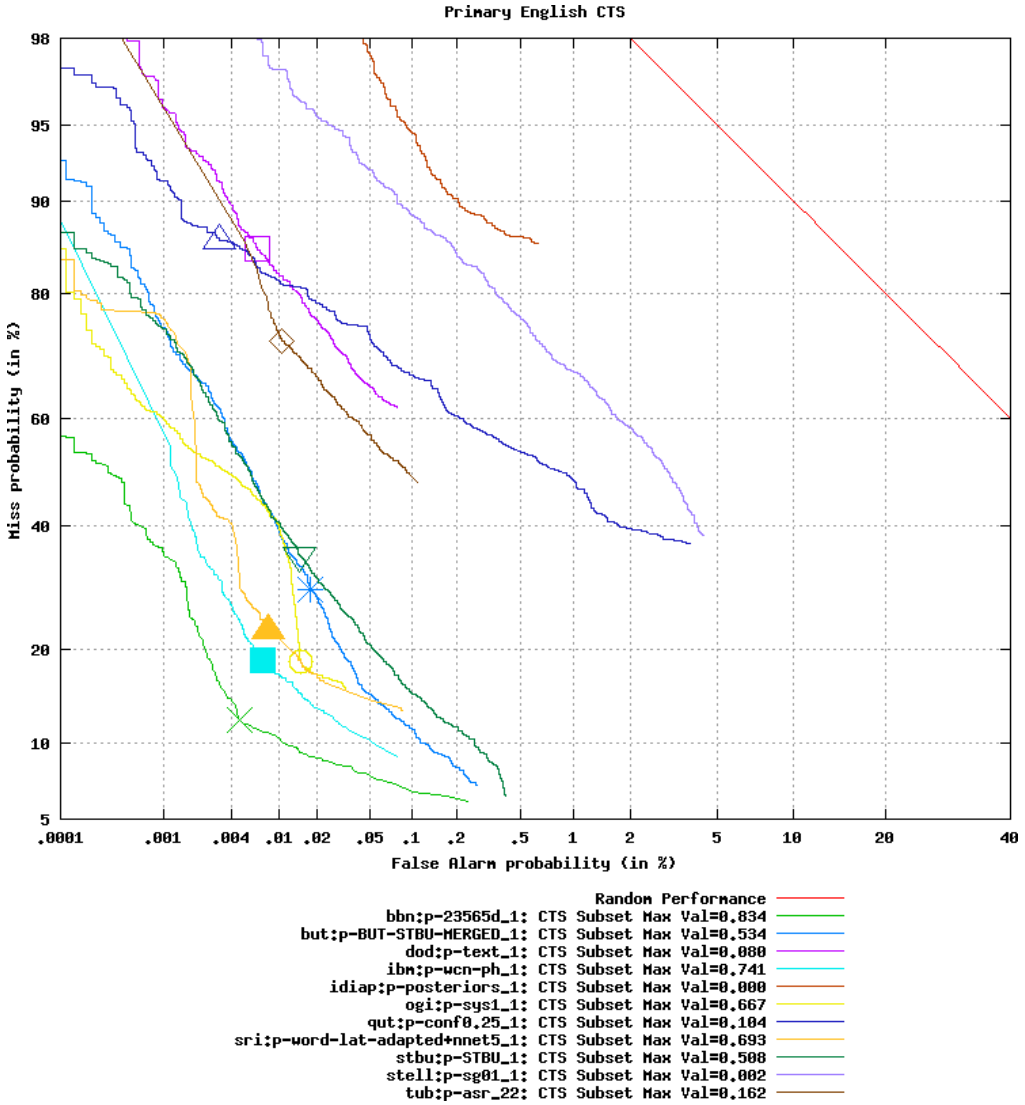
Site	ATWV	MTWV
BBN	0.8335	0.8336
BUT	0.5235	0.5344
DOD	-0.4070	0.0801
IBM	0.7392	0.7408
IDIAP	-6.1927	0.0000
OGI	0.6470	0.6671
QUT	0.0873	0.1044
SRI	0.6652	0.6934
STBU	0.2152	0.5081
STELL	-0.0002	0.0022
TUB	0.1598	0.1617



ATWV: Actual Term Weighted Value
MTWV: Maximum Term Weighted Value

Performance - CTS

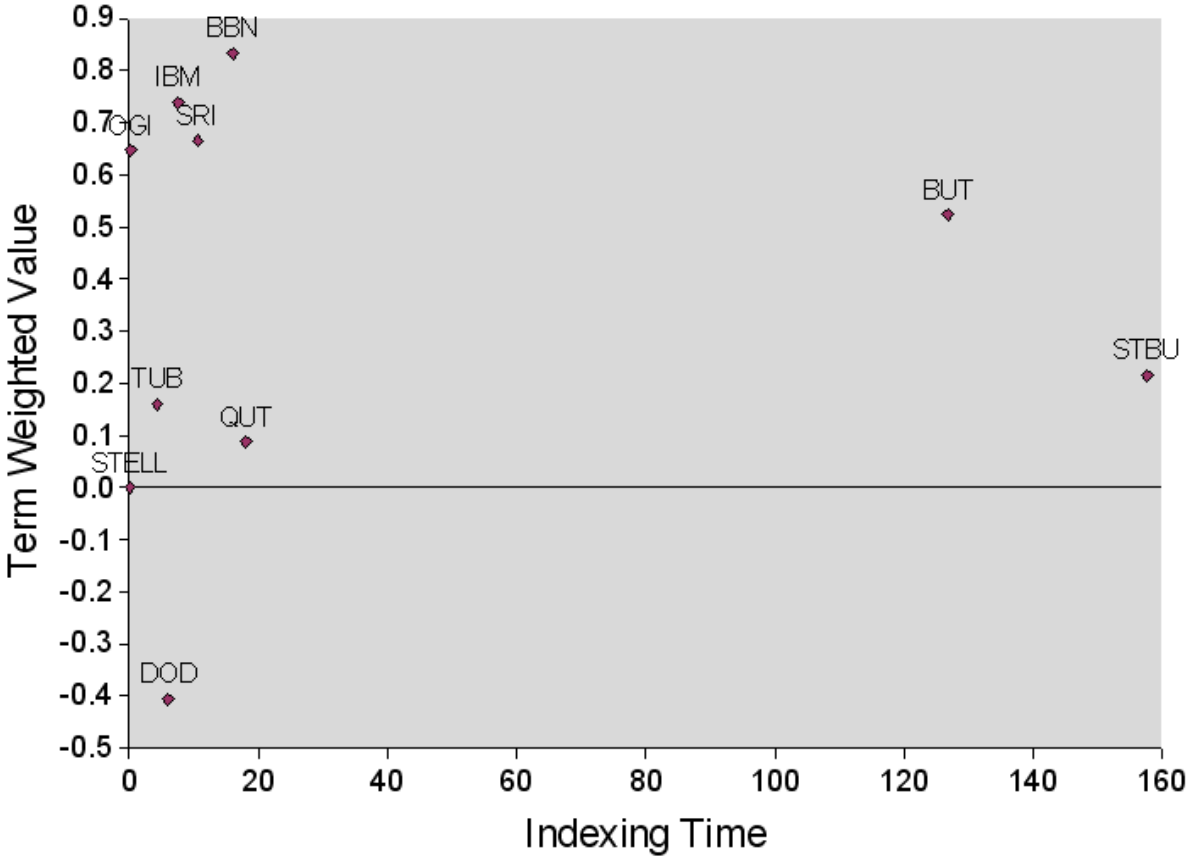
DET Curve



Indexing Time vs. ATWV

Site	Indexing Time	ATWV
BBN	16.1087	0.8335
BUT	126.7596	0.5235
DOD	6.0232	-0.4070
IBM	7.5627	0.7392
IDIAP	0.2859	-6.1927
OGI	0.2891	0.6470
QUT	18.0884	0.0873
SRI	10.6555	0.6652
STBU	157.6593	0.2152
STELL	0.1680	-0.0002
TUB	4.3897	0.1598

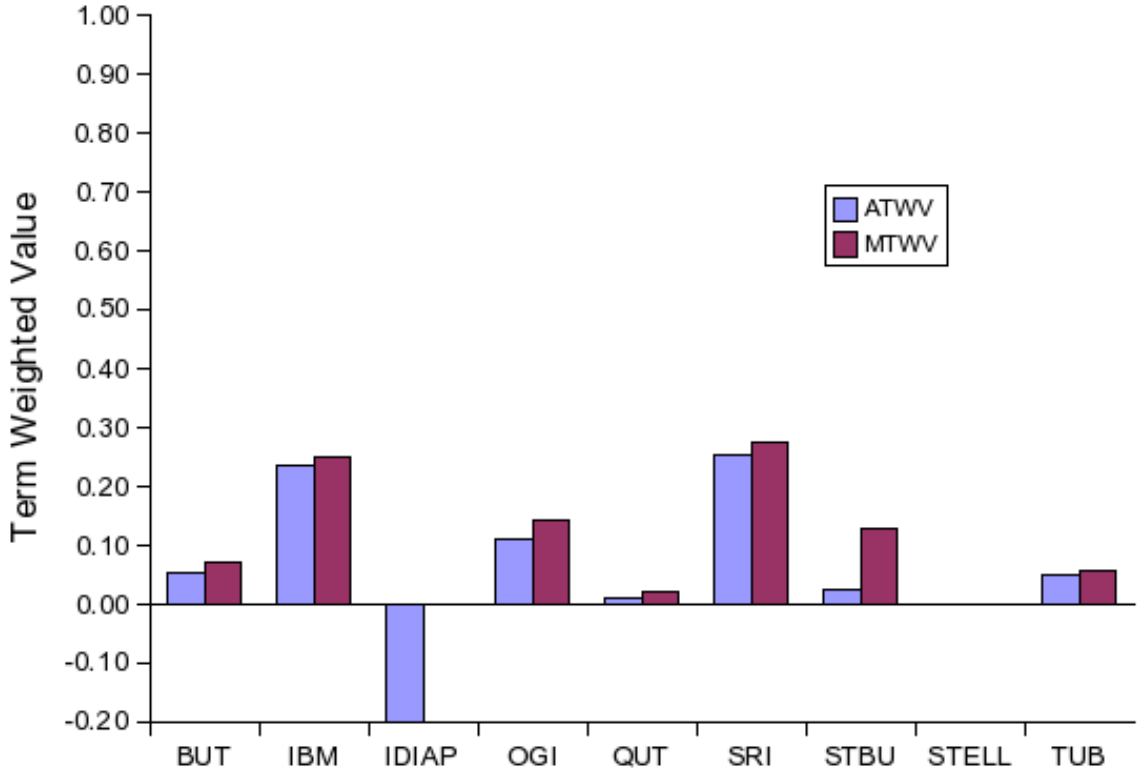
Indexing Time Unit:
Processing Hours per Hours of Speech



Actual and Maximum Term Weighted Value

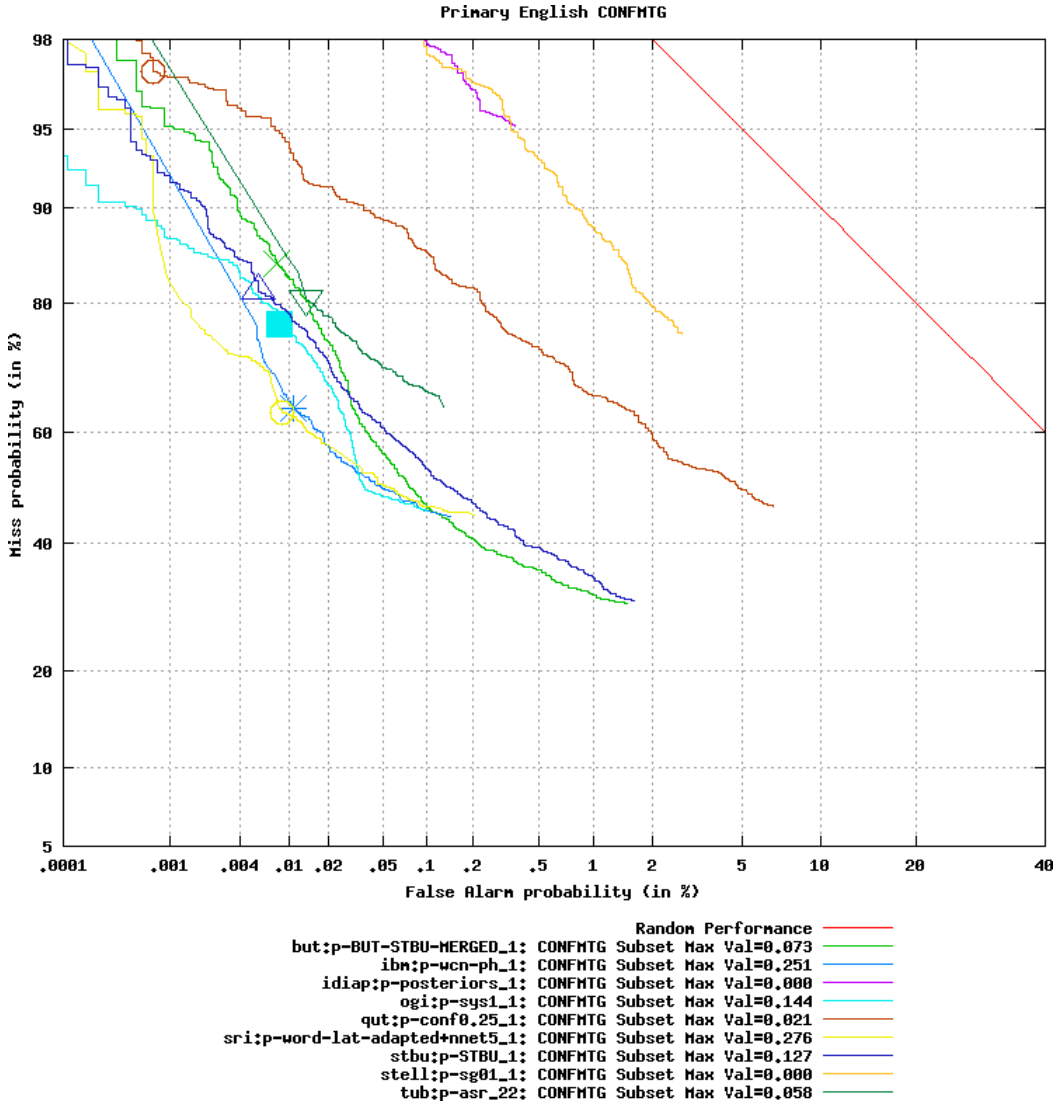
Site	ATWV	MTWV
BUT	0.0549	0.0731
IBM	0.2365	0.2508
IDIAP	-3.6170	0.0000
OGI	0.1103	0.1437
QUT	0.0109	0.0213
SRI	0.2553	0.2765
STBU	0.0253	0.1273
STELL	0.0000	0.0000
TUB	0.0500	0.0578

ATWV: Actual Term Weighted Value
MTWV: Maximum Term Weighted Value



Performance - CONFMTG

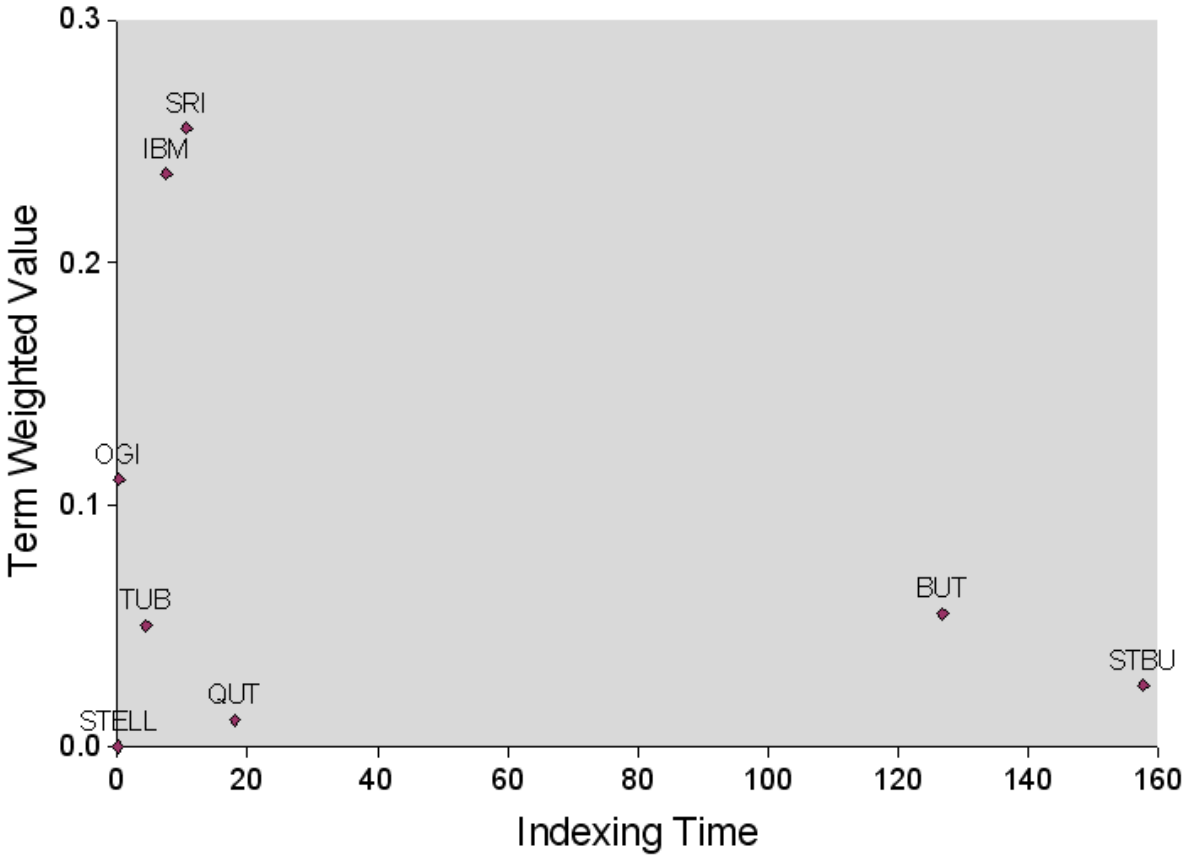
DET Curve



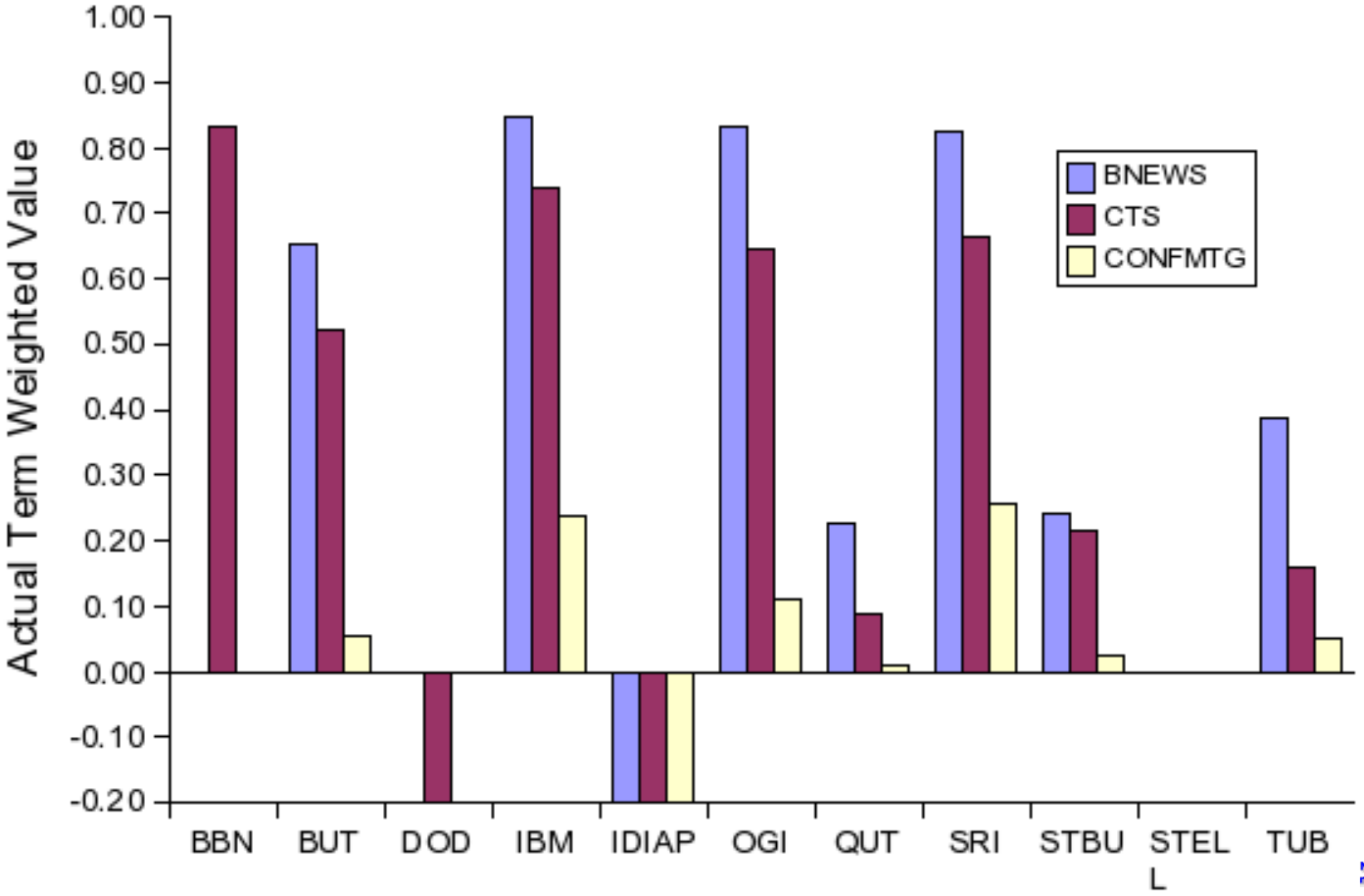
Indexing Time vs. ATWV

Site	Indexing Time	ATWV
BUT	126.7596	0.0549
IBM	7.5627	0.2365
IDIAP	0.2859	-3.6170
OGI	0.2891	0.1103
QUT	18.0884	0.0109
SRI	10.6555	0.2553
STBU	157.6593	0.0253
STELL	0.1680	0.0000
TUB	4.3897	0.0500

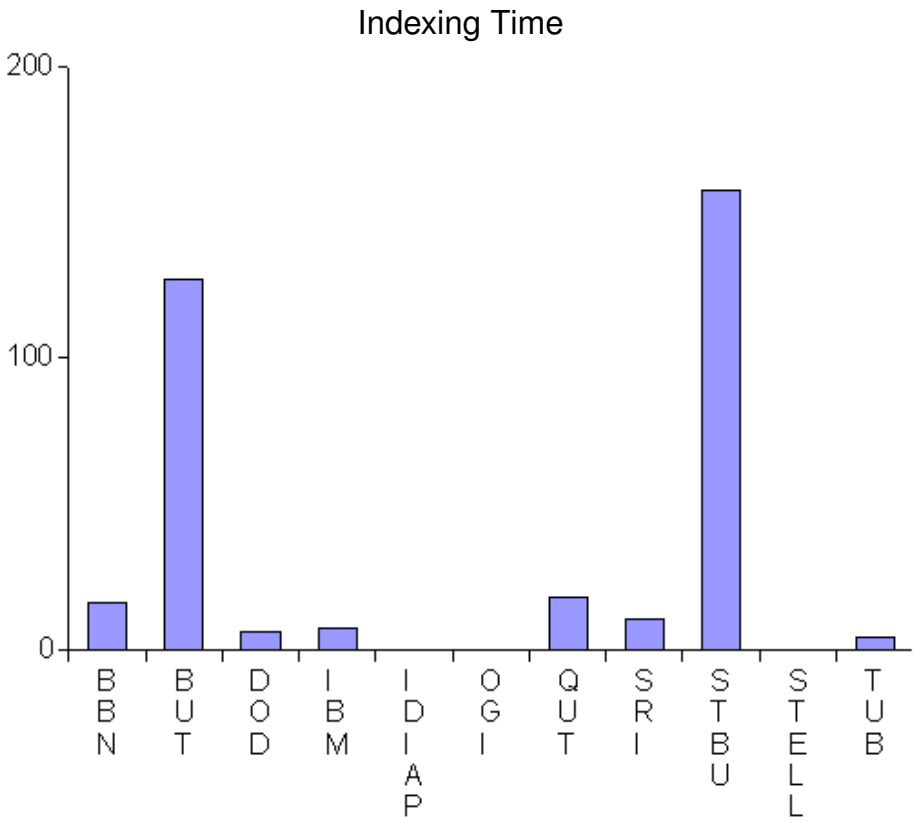
Indexing Time Unit:
Processing Hours per Hours of Speech



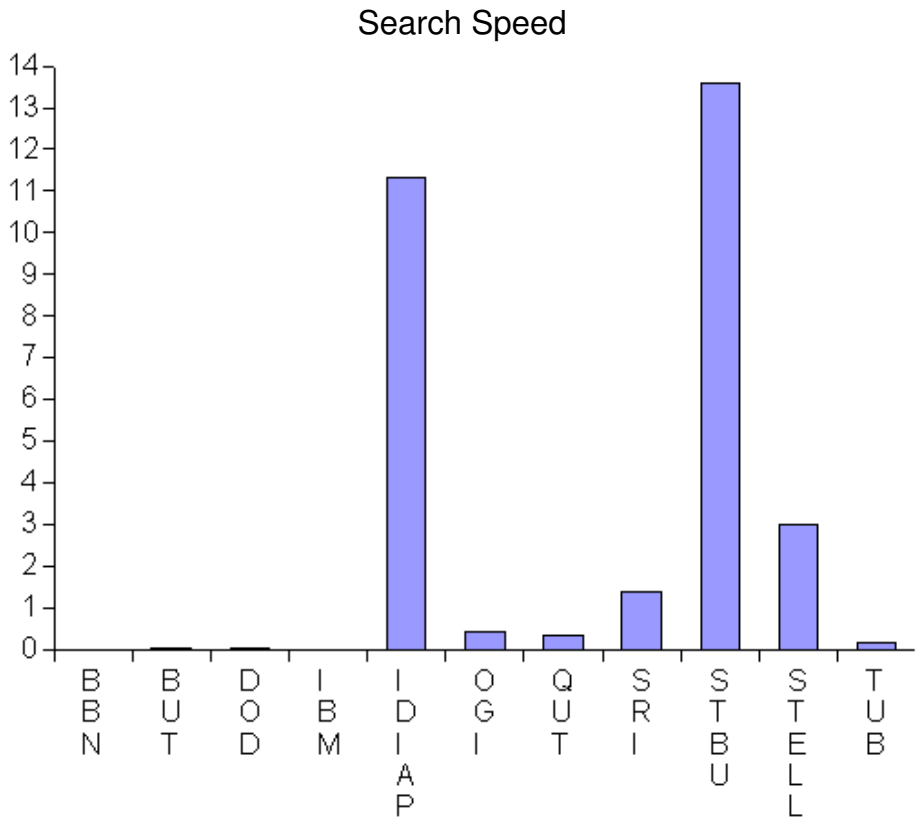
Performance - ATWV Overview



Indexing Time & Search Speed

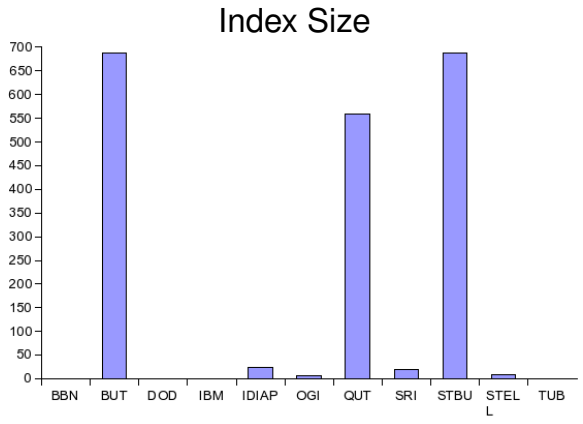


Unit: Processing Hours per Hours of Speech

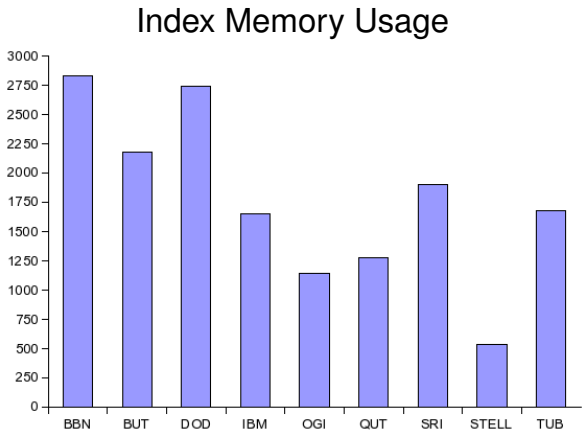


Unit: Processing seconds per Hours of Speech

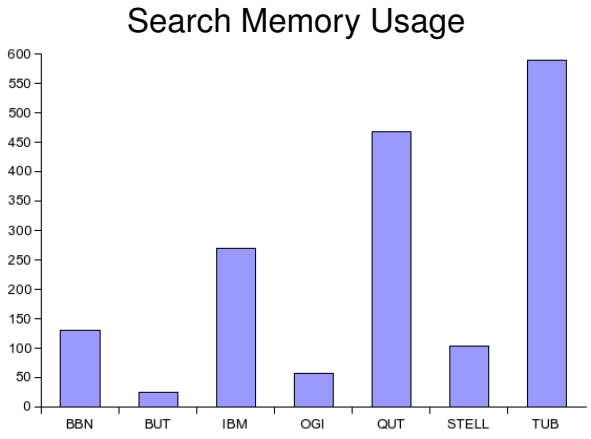
Index Size & Memory Usages



Unit: MegaBytes per Processing Hours



Unit: MegaBytes



Unit: MegaBytes

Outline

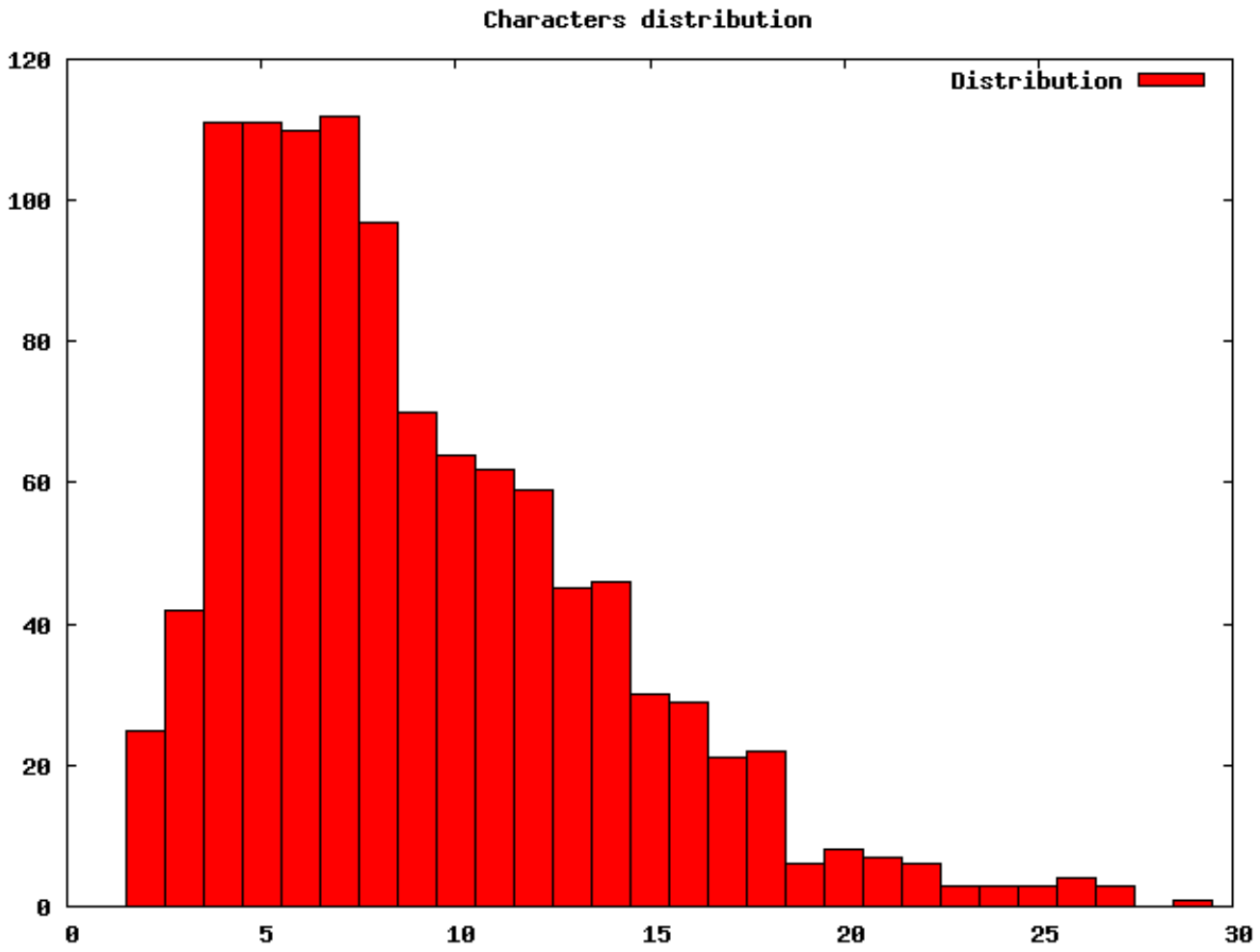
1. Participants
   - Number of Submissions
2. Corpus Information
   - Terms Richness
3. Primary Results
   - Performance - BNEWS
   - Performance - CTS
   - Performance - CONFMTG
   - Performance - ATWV Overview
   - Processing Ressources
4. **Conditional Analysis**
   - Analyses
   - Distribution
   - N-gram
   - Syllable Quantile
   - Occurrence in all sources

List of Analyses

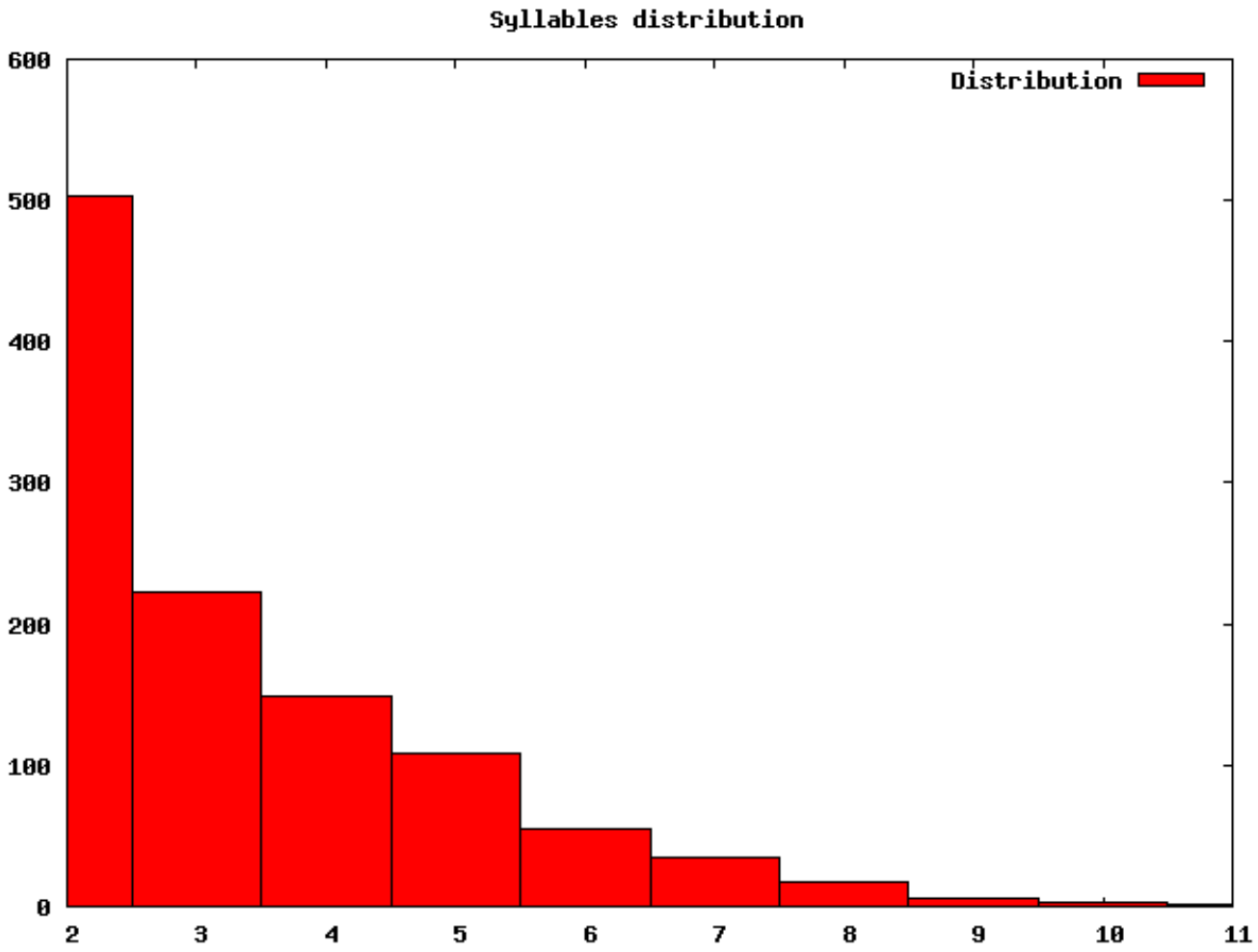
Term Sub-sets

- Characters length quantile
- Average duration of reference occurrence
- Number of words
- Occurrence in all the sources
- Out and In of Vocabulary
- Syllables quantile

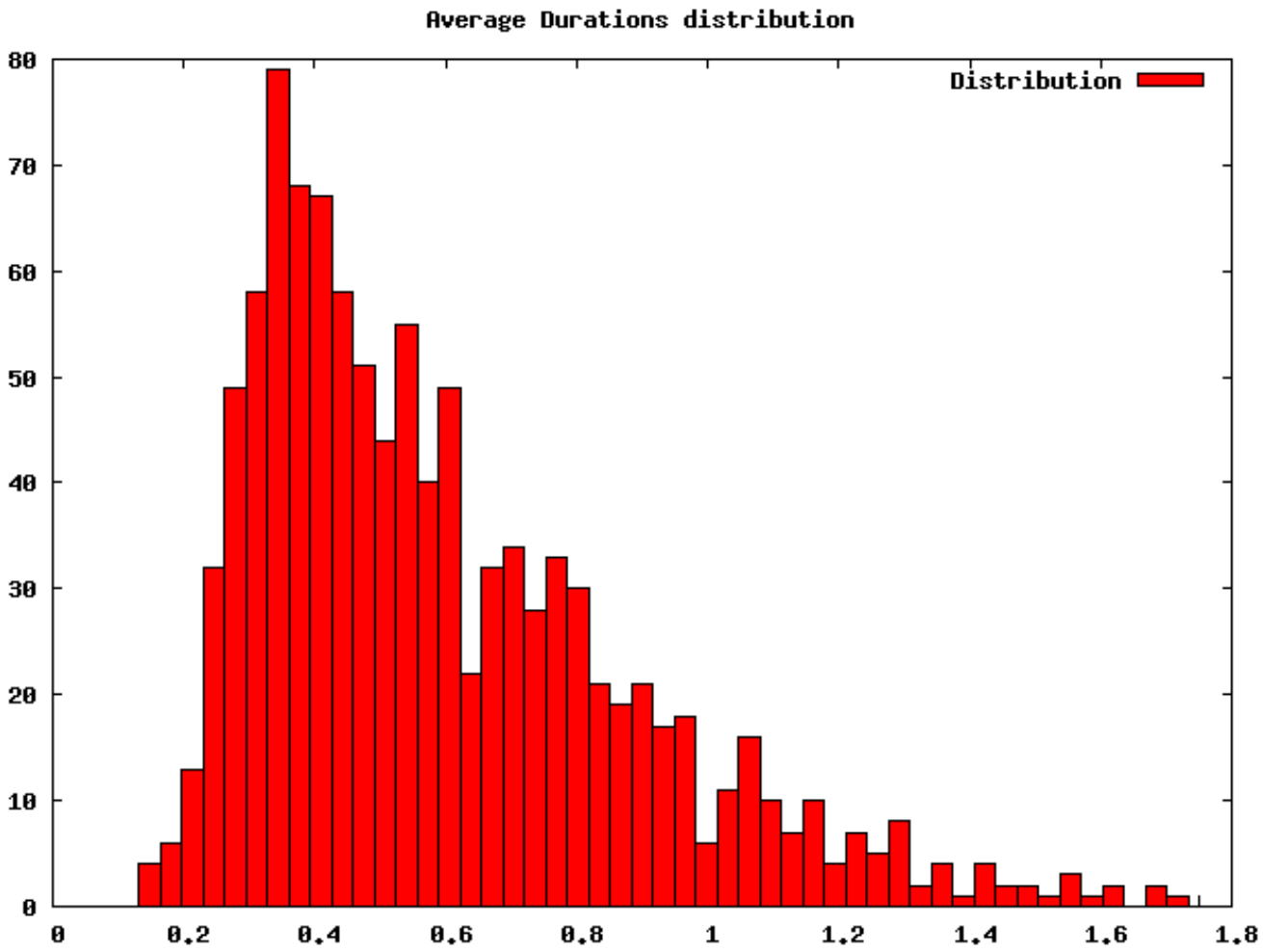
Characters


Characters distribution
Distribution
120
100
80
60
40
20
0
0
5
10
15
20
25
30

Syllables


Syllables distribution
Distribution
600
500
400
300
200
100
0
2
3
4
5
6
7
8
9
10
11

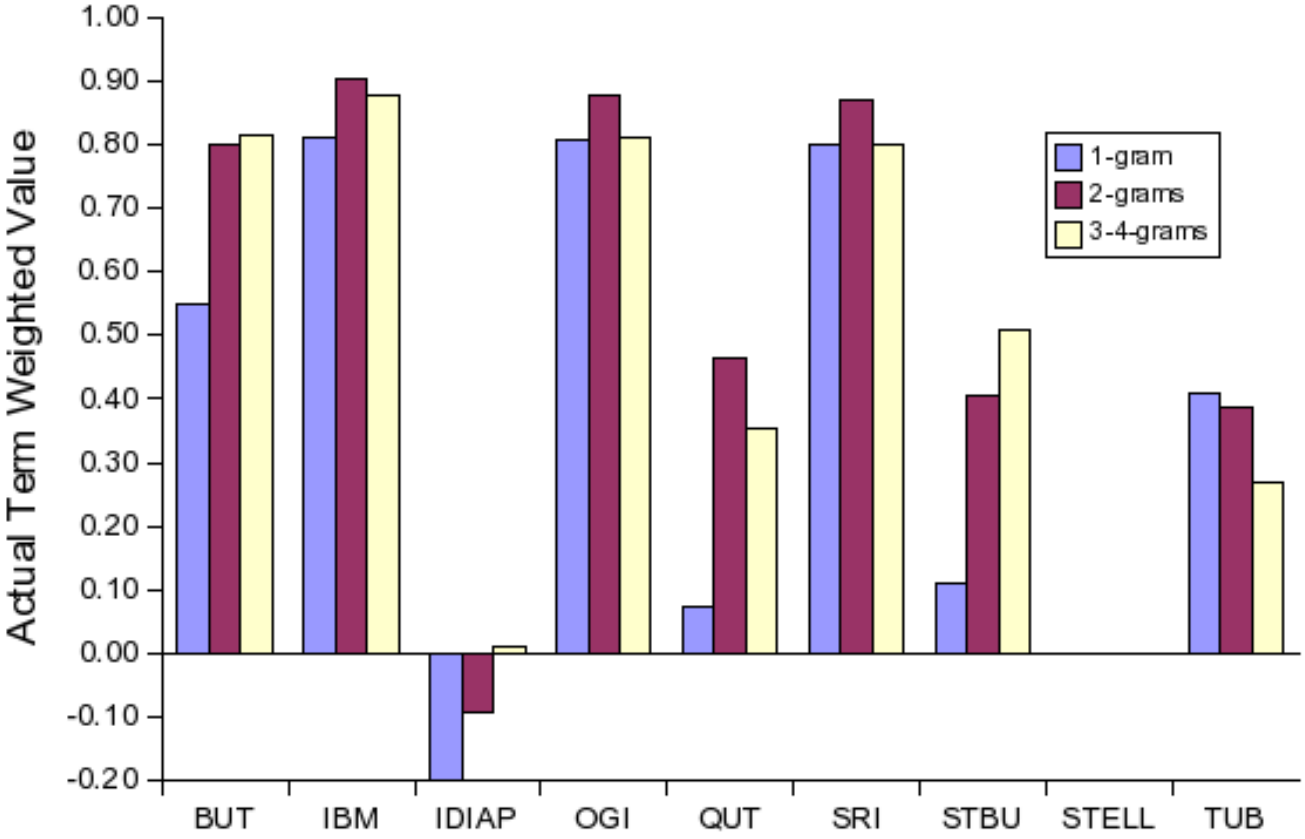
Duration


Average Durations distribution
Distribution
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8

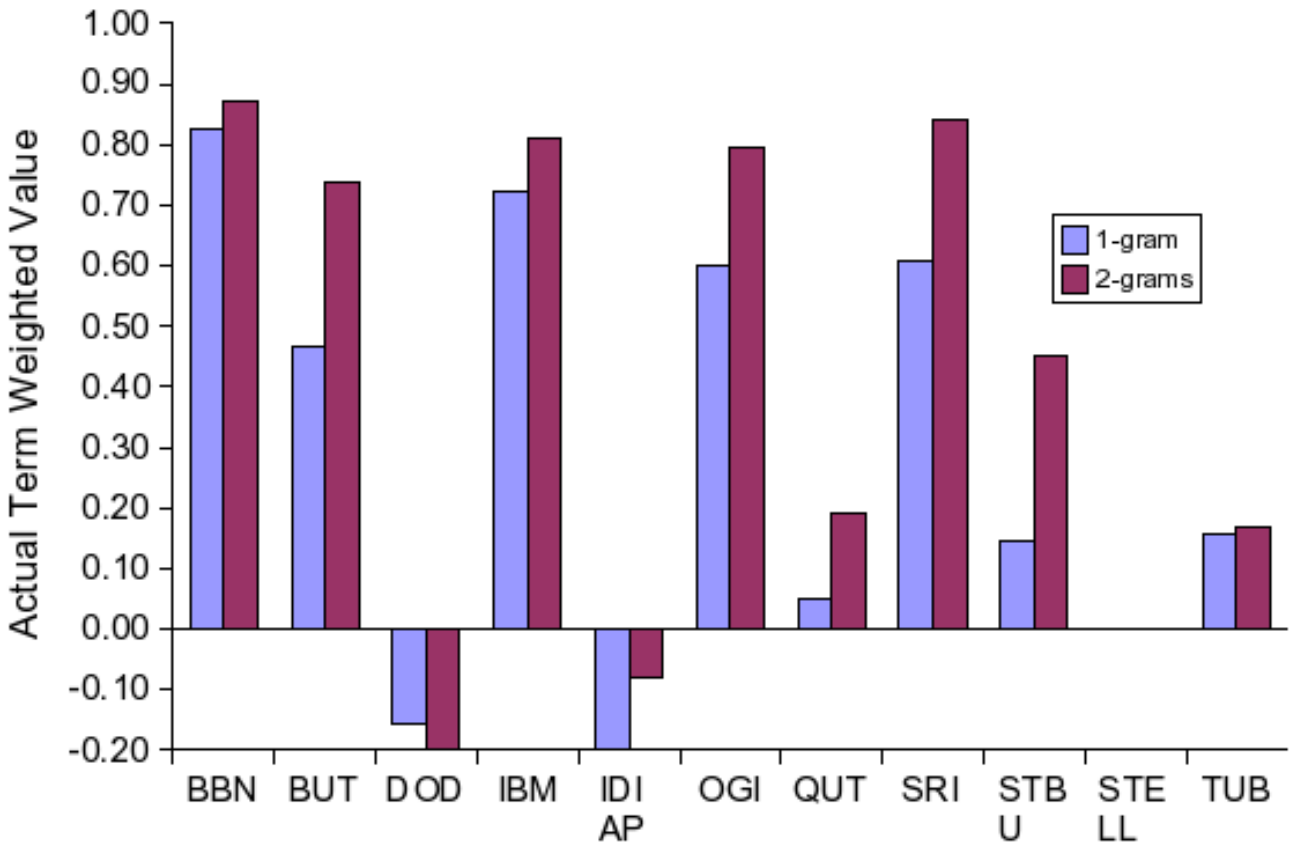
Distribution

Source type	N-Gram	Terms		Ocurrences	
		Count	%	Count	%
BNEWS	1-gram	523	58.24	4187	85.57
	2-grams	298	33.18	608	12.43
	3-4-grams	77	8.57	98	2.00
CTS	1-gram	314	76.40	5674	96.89
	2-grams	83	20.20	166	2.83
	3-4-grams	14	3.41	16	0.27
CONFMTG	1-gram	212	87.97	3614	98.42
	2-grams	27	11.20	55	1.50
	3-4-grams	2	0.83	3	0.08

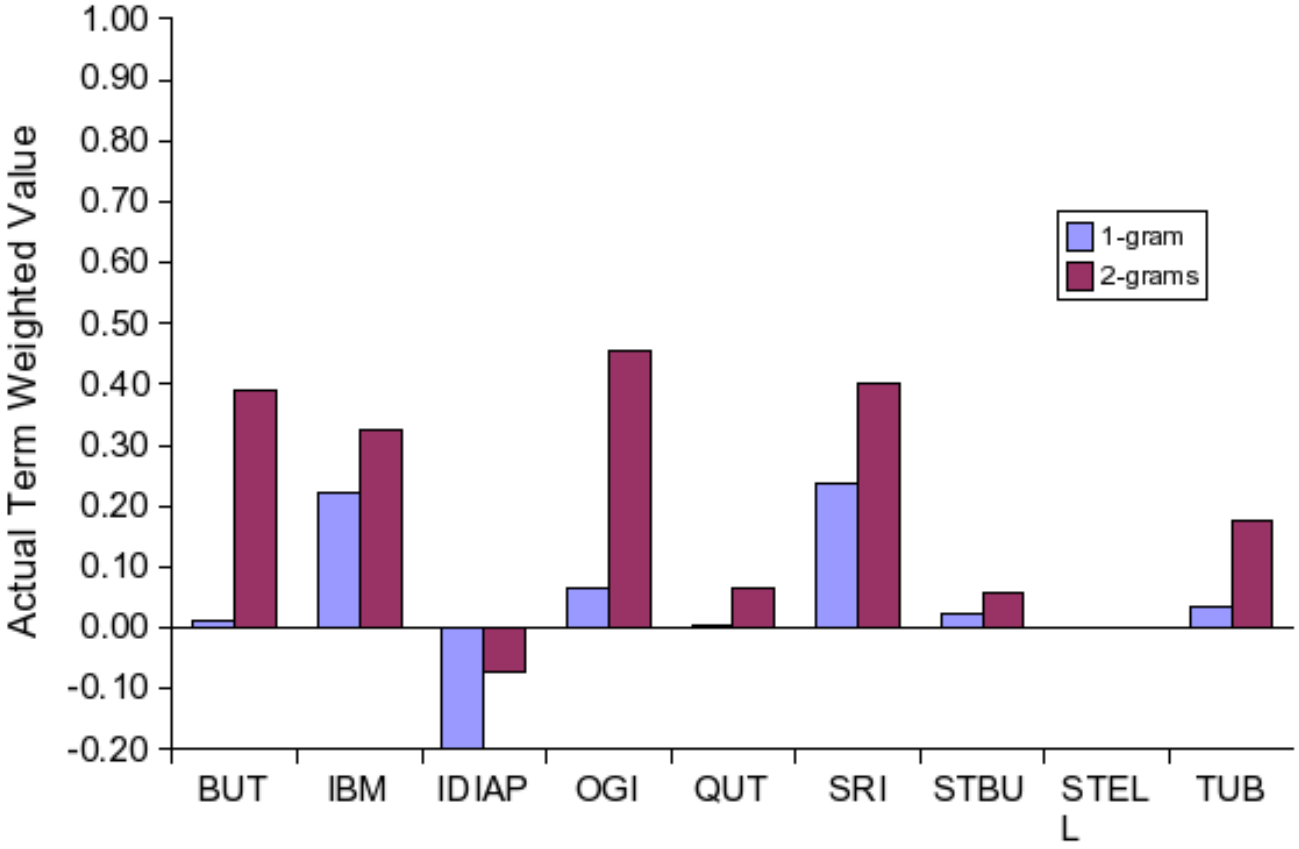
BNEWS - Actual Term Weighted Values



CTS - Term Weighted Values



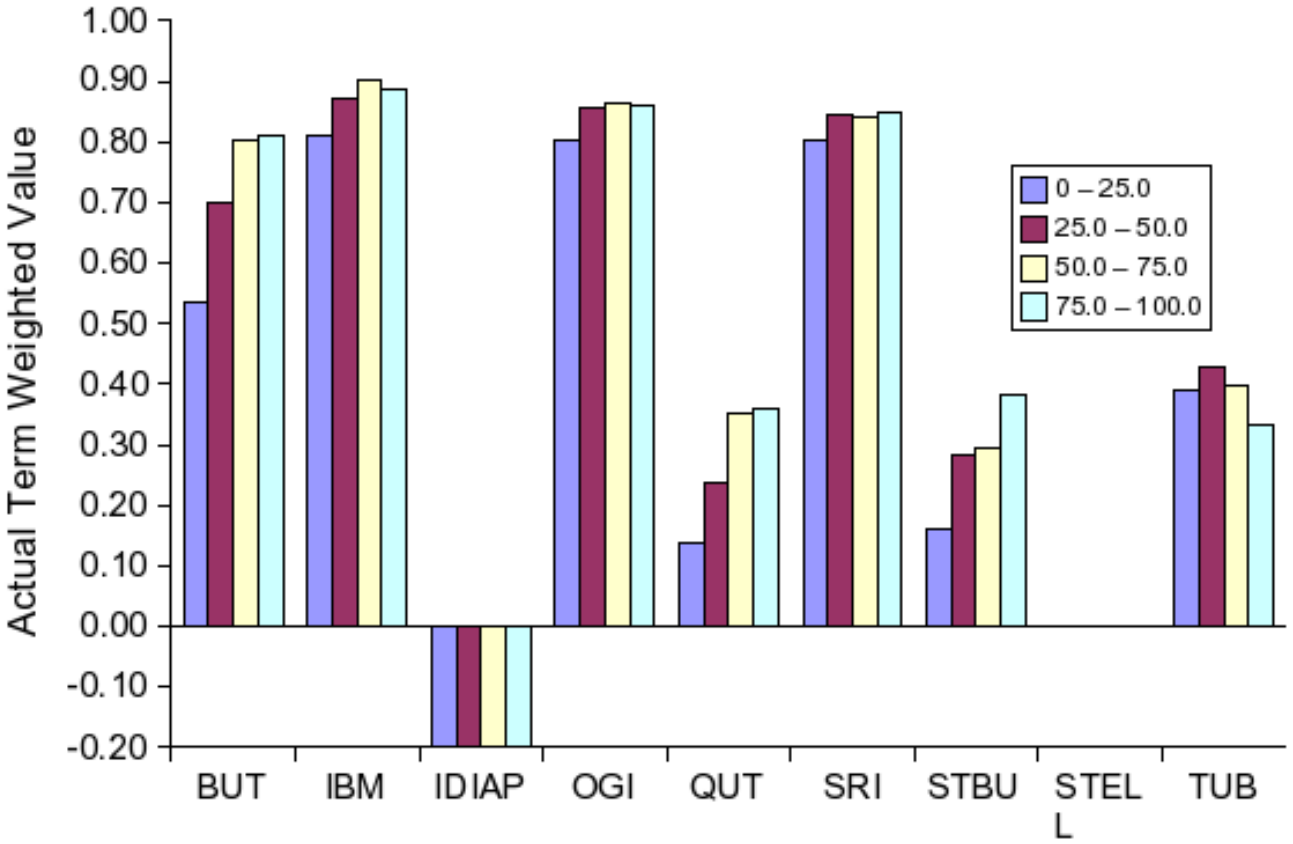
CONFMTG - Term Weighted Values



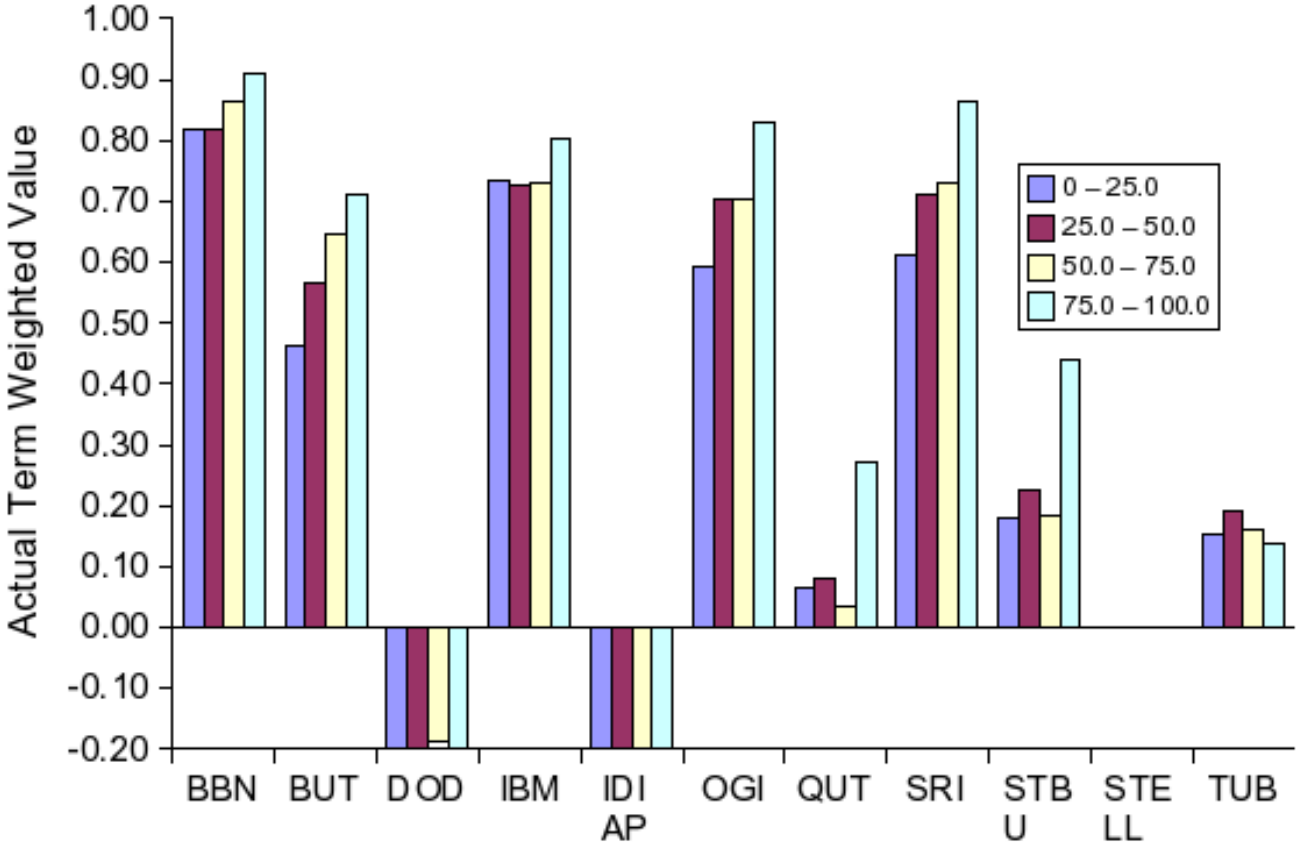
Distribution

Source type	N-Gram	Terms		Ocurrences	
		Count	%	Count	%
BNEWS	0 - 25.0	439	48.89	3722	76.07
	25.0 - 50.0	181	20.16	668	13.65
	50.0 - 75.0	116	12.92	295	6.03
	75.0 - 100.0	162	18.04	208	4.25
CTS	0 - 25.0	260	63.26	5392	92.08
	25.0 - 50.0	70	17.03	256	4.37
	50.0 - 75.0	37	9.00	125	2.13
	75.0 - 100.0	44	10.71	83	1.42
CONFMTG	0 - 25.0	178	73.86	3429	93.38
	25.0 - 50.0	33	13.69	162	4.41
	50.0 - 75.0	17	7.05	50	1.36
	75.0 - 100.0	13	5.39	31	0.84

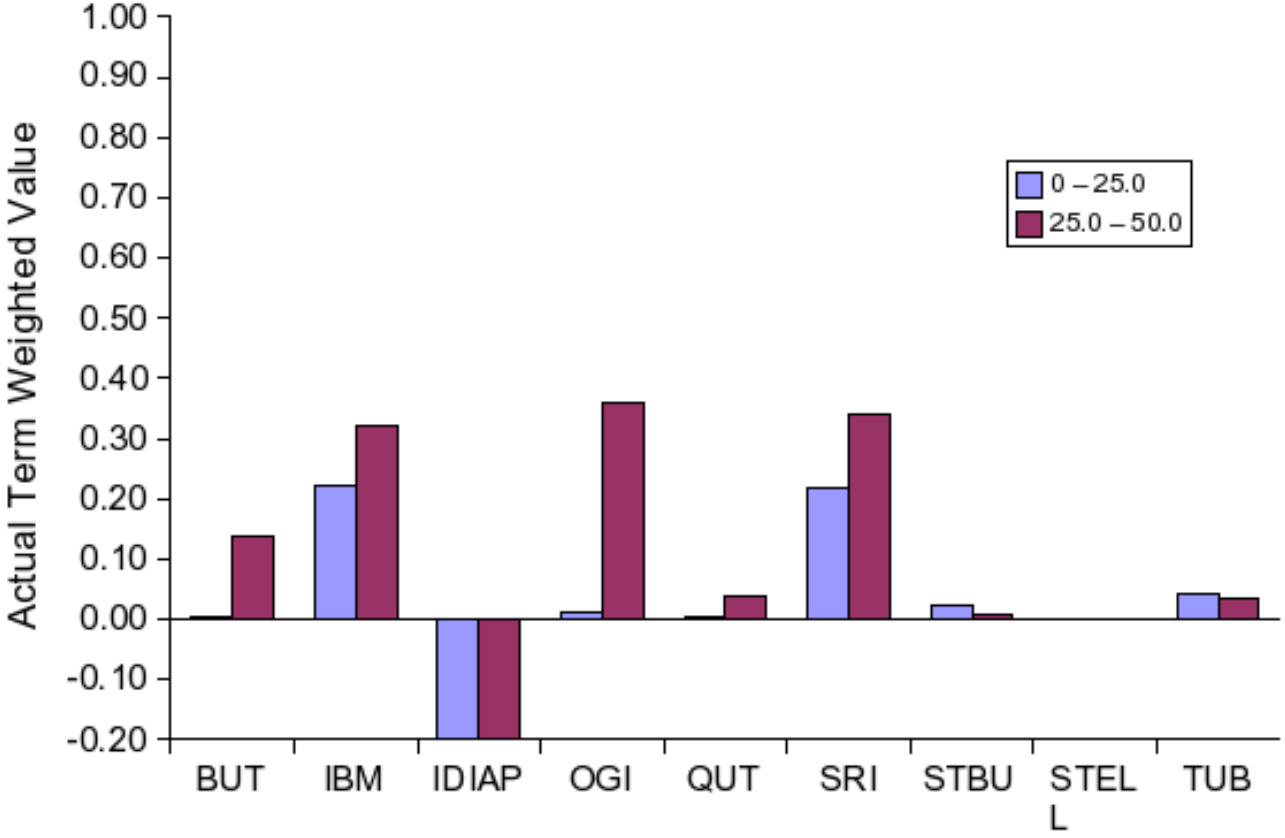
BNEWS - Term Weighted Values



CTS - Term Weighted Values



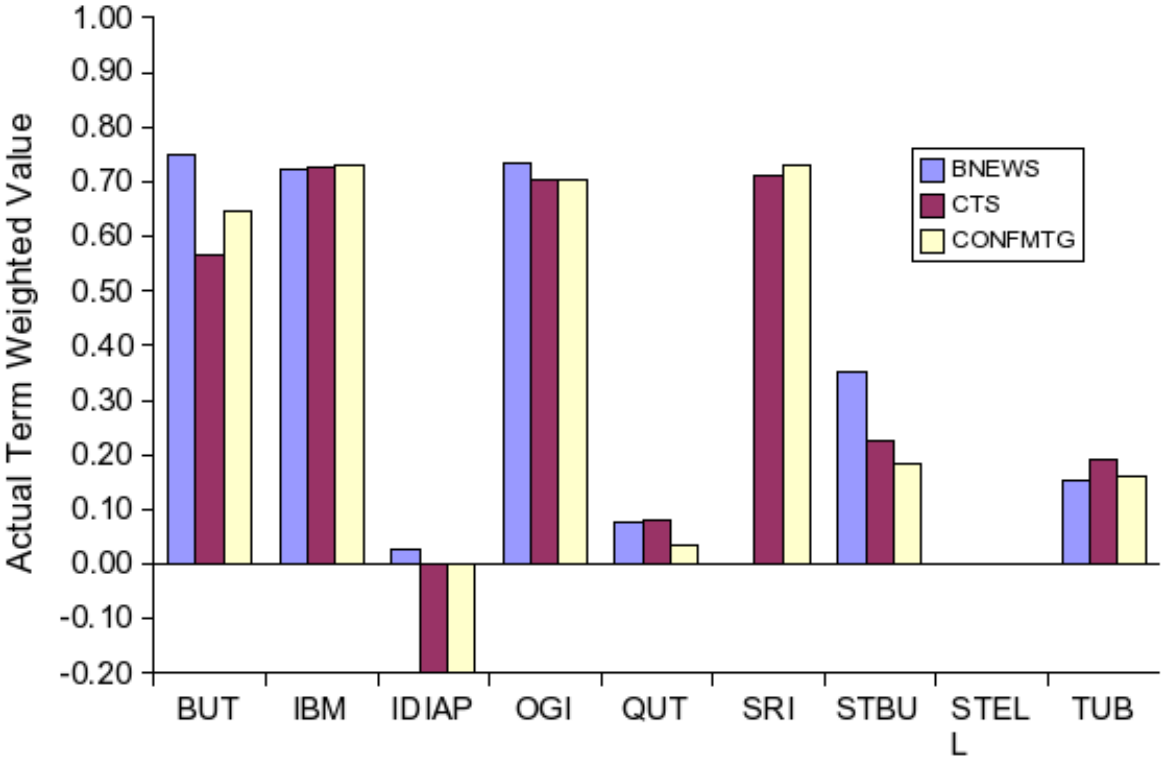
CONFMTG - Term Weighted Values



Occurrence in all sources

 # Terms: 160

Source type	# Occurrences
BNEWS	2452
CTS	5096
CONFMTG	3433



Summary

Winner

- BUT well... I think

Losers

- IDIAP
- STELL